Database Resources of the NCBI

Peter S. Cooper

email: cooper@ncbi.nlm.nih.gov
phone: 301-435-5951

updated 8/9/01

Online Resources

Lecture Resources:

Field Guide Home:	http://www.ncbi.nlm.nih.gov/Class/FieldGuide/
Power Point Slides:	ftp://ftp.ncbi.nih.gov/pub/cooper/
Problem Set:	http://www.ncbi.nlm.nih.gov/Class/FieldGuide/problem_set.html
Course Links:	http://www.ncbi.nlm.nih.gov/Class/FieldGuide/links.html

General:

NCBI Homepage:	http://www.ncbi.nlm.nih.gov
Site Map:	http://www.ncbi.nlm.nih.gov/Sitemap/index.html
NCBI News:	http://www.ncbi.nlm.nih.gov/About/newsletter.html

GenBank:

GenBank Release Notes:	ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt
dbEST:	http://www.ncbi.nlm.nih.gov/dbEST/index.html
dbSTS:	http://www.ncbi.nlm.nih.gov/dbSTS/index.html
dbGSS:	http://www.ncbi.nlm.nih.gov/dbGSS/index.html

Collaborating Nucleotide Databases:

EMBL:	http://www.ebi.ac.uk/
DDBJ:	http://www.ddbj.nig.ac.jp/

Entrez:

Entrez:	http://www.ncbi.nlm.nih.gov/Entrez/
Network Entrez:	ftp://ncbi.nlm.nih.gov/entrez/CURRENT/

BLAST:

BLAST Main Page:	http://www.ncbi.nlm.nih.gov/BLAST/
Stephen Altschul's Lectures on BLAST statistics:	http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html
Frequently Asked Questions:	http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html

BLAST tutorials: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

BLAST Clients, Executables and Databases:	ftp://ncbi.nlm.nih.gov/blast/
NCBI Source Code:	ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools/

NCBI Structures:

Structure Homepage:	http://www.ncbi.nlm.nih.gov/Structure/
Cn3D tutorial:	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dtut.html

Genomic Resources:

Entrez Genomes:	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Human Genome Resources:	http://www.ncbi.nlm.nih.gov/genome/guide/
Map Viewer	http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch/
LocusLink:	http://www.ncbi.nlm.nih.gov/LocusLink/
UniGene:	http://www.ncbi.nlm.nih.gov/UniGene/
CGAP:	http://cgap.nci.nih.gov/

Other Databases:

SWISS-PROT	http://expasy.cbr.nrc.ca/sprot/
PIR	http://pir.georgetown.edu/pirwww/pirhome.shtml
PDB	http://www.rcsb.org/pdb/
PRF	http://www.prf.or.jp/en/

Email addresses:

Humans:

General Help	info@ncbi.nlm.nih.gov
Updates	update@ncbi.nlm.nih.gov
BLAST	blast-help@ncbi.nlm.nih.gov
Sequin Submissions	gb-sub@ncbi.nlm.nih.gov
Batch Submissions	batch-sub@ncbi.nlm.nih.gov

Servers:

BLAST Server:	blast@ncbi.nlm.nih.gov
Query Server:	query@ncbi.nlm.nih.gov

Literature Reference List

General

Baxevanis, A. and Ouellette, B.F.F., eds. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Second Edition.* New York: John Wiley & Sons. 2001. ISBN: 0-471-38391-0

Gibas, C. and Jambeck, P. *Developing Bioinformatics Computer Skills.* Sebastopol: O'Reilly and Associates. 2001. ISBN: 1-56592-664-1

Wheeler DL, et al. 2001. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **29**(1):11-16. PMID: 11125038

GenBank Database

Benson DA, et al. 2000. Genbank. *Nucleic Acids Res.* **28**(1):15-18. PMID: 10592170

Ouellette BF, Boguski MS.1997. Database divisions and homology search files: a guide for the perplexed. *Genome Res*. **7**(10):952-5. PMID: 9331365

BLAST

Altschul SF, et al. 1990. Basic local alignment search tool. *J Mol Biol.* **215**(3):403-10. PMID: 2231712.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17):3389-402. PMID: 9254694.

Altschul SF, et al. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci.* **23**(11):444-7. PMID: 9852764.

Schaffer AA, et al. 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics.***15**(12):1000-11. PMID: 10745990.

Schaffer AA, et al. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**(14):2994-3005

Tatusova TA, et al. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.* **174**(2):247-50. PMID: 10339815.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* **7**(1-2):203-14.PMID: 10890397; UI: 20346451

Zhang Z, et al. 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26**(17):3986-90. PMID: 9705509.

MMDB and Structures

Hogue CW. 1997. Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem Sci*. **22**(8):314-6. PMID: 9270306.

Wang Y, et al. 2000. MMDB: 3D structure data in Entrez. *Nucleic Acids Res*. **28**(1):243-245. PMID: 10592236

Specialized Genomic Resources

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921

Jang W, et al. 1999. Making effective use of human genomic sequence data. *Trends Genet*. **15**(7):284-6. PMID: 10390628.

Pruitt KD Maglott DR,. 2001 RefSeq and LocusLink: NCBI gene centered resources. *Nucleic Acids Res.* **29**(1):137-140. PMID: 11125071.

Sherry, S.T., et al. 2000. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**(1):308-311. PMID: 11125122

Tatusov RL, et al. 2001 The COG database: new developments in the phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**(1):22-28. PMID: 11125040

Wolfsberg, T. et al. 2001. Guide to the draft human genome. *Nature* **409**, 824 - 826.

Sequence Database Notes:

Primary Sequence Databases

Primary sequence databases contain original reports of biological sequences from the investigators that determined them. Typically there are no additional data or interpretation added by the maintainers of these databases. These collections are archives of sequence information in that they contain records that may be unmodified or updated from the time they were submitted. A second consequence of this archival nature is redundancy in the data; there may be many examples of the same target sequence in the database submitted by different researchers.

The earliest examples of primary sequence databases are protein sequence collections. The Protein Information Resource (PIR) at Georgetown University is a direct descendant of one of the early protein sequence collections. As DNA cloning and sequencing technology improved, the number of nucleotide sequences available quickly surpassed directly determined amino acid sequences. Today by far the largest primary sequence databases are nucleic acid sequence databases. In fact the majority of protein sequences available today are based on conceptual translations of the coding regions on DNA sequences.

The International DNA Sequence Database Collaboration

The most important primary DNA sequence databases today are the three members of the international DNA sequence collaboration: GenBank, the European Molecular Biology Laboratory (EMBL) database and the DNA Database of Japan (DDBJ). All three of these databases accept direct submissions of sequence data and are products of government-sponsored institutions in their respective countries. GenBank is produced and maintained by the National Center for Biotechnology Information at the National Institutes of Health in the United States, the EMBL database by the European Bioinformatics Institute in the United Kingdom and DDBJ by the Center for Information Biology of the National Institute of Genetics in Japan. All of these entities maintain a presence on the World Wide Web that includes browser-based access to data and tools for sequence analysis. The scope of these centers includes more than just their primary database product. All are active centers of computational biology and bioinformatics research and produce other data products including many important derivative databases.

The discussion here will focus on GenBank. Most of most of what is said will apply to the other two databases as well.

GenBank

The GenBank database has its origins in the dim past when it was produced in bound volumes. As the number of sequences increased and computer technology advanced, the database was made available on CD-ROM and came with software for accessing the data. The CD-ROM version was discontinued in 1997 when the number of CDs required became prohibitive. Right now GenBank is available through the Internet on the NCBI ftp site (URL: ftp://ftp.ncbi.nlm.nih.gov/genbank/). On the NCBI ftp server the database is made available as full data releases every two months in the even numbered months of the year. Between releases daily updates are provided. For each release, important information including release statistics is in the Release Notes (URL: ftp://ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt). The current release contains over 12 billion bases and more than 12 million sequences from over 80,000 species.

Data files and GenBank Division Codes

On the ftp site, the GenBank data are divided into series of sequence files. Originally each of these sequence files corresponded to one GenBank division. The GenBank divisions are identified by a three-letter division code, for example the BCT (bacterial) division or the PRI (primate) division. The days of one file per division are long past now; the EST division is split into more than 100 files. However all of the more than 11 million sequences in GenBank are still separated into a handful of divisions. A discussion of GenBank divisions is helpful in classifying the kinds of data in GenBank and is also useful in searching the data on the NCBI web site using the Entrez system. For the purposes of this discussion we recognize two kinds of GenBank divisions, traditional and special or bulk sequence divisions.

Traditional Divisions

The traditional GenBank divisions contain sequences that are determined to a high degree of accuracy (1 error in 10,000,) and often have extensive annotation about the biology or features of the sequence. At first glance these traditional divisions appear taxonomic in nature. Closer inspection shows the overriding purpose in establishing them initially was to create single files of reasonable size. Taxa were spilt or lumped to accomplish this. For example, the primate (PRI) and rodent (ROD) sequences were separated from the rest of the mammalian sequences (MAM) because there were a large number of primate and rodent sequences. On the other hand the fungal and plant sequences were lumped into the PLN division because originally there were fewer of these. Sequence data can be submitted to these divisions through a web based form called BankIt (URL: http://www.ncbi.nlm.nih.gov/BankIt/) more complex submissions can be prepared using the NCBI standalone tool Sequin (URL: http://www.ncbi.nlm.nih.gov/Sequin/index.html)

Special Divisions

With changes in DNA sequencing technology and strategies, a number of special GenBank divisions were established. These are not based on the source organism of the sequence but are based on the technique used to generate the sequence or the intended use of the sequence. A unifying characteristic of these divisions is that they tend to be submitted in large batches by single projects. The GSS, EST and STS sequences also exist in NCBI databases apart from GenBank: dbGSS, dbEST and dbSTS. The format of the records within these databases is quite different than that used in GenBank. Submissions to these divisions are handled through different procedures and different staff than traditional submissions.

First pass sequence divisions

The expressed sequence tag (EST) division and the genome survey sequence division (GSS) were established to hold first pass single read sequences that have little or no annotation. Because these data are single sequence reads, the amount of sequence in each record is limited, and is likely contain sequencing errors including frame shifts and base miscalls.

EST division

The EST division holds automatically generated partial cDNA sequences. These sequences are derived from arrayed cDNA libraries. For each clone in the library, only a single read is obtained from each end of the insert using the standard sequencing primers. Thus there can be two sequences in GenBank for each clone. In the case of directionally cloned libraries these will be the 5'end of the cDNA and the 3'end. Robots are often used to automate the process of sequencing these clones. Thus, large numbers of clones can be partially sequenced very rapidly using this strategy. Nearly comprehensive sets of EST data are available for a number of organisms. In fact the EST division is the largest division of GenBank. Although largely unannotated and error prone, these data provide a rich source of information about the expressed sequences in a particular cell type tissue or ultimately the organism. The EST data are an important resource gene discovery and gene expression data. At the NCBI, the EST data have been used to generate a derivative database; UniGene that attempts to organize these data into gene based clusters.

GSS division

The GSS division contains data that are the genomic equivalent of the EST data. That is first pass single reads of genomic clones. The bulk of the data in the GSS division are derived from bacterial artificial chromosome (BAC) libraries. BACs are the large insert genomic clones that are used in complex genome projects like the human genome project. This is explained in more detail below when we describe the HTG division. Sequencing centers will produce preliminary reads for these clones sometimes as a prelude to producing more complete sequences. These surveys go in the GSS division. Another related category of sequence comes from the extension of sequencing primers onto the

insert of the clone. These so called BAC end sequences are used to identify overlapping clones and creating tiling paths for assembling large genomic contigs. The GSS division also contains whole genome shotgun sequencing reads for some organisms, most notably certain protozoan parasites. These GSS sequences are important resources for genomic sequence for these organisms even in this unassembled form.

The High Throughput Genome (HTG) Sequence Division

Many of the large-scale genome-sequencing projects rely on a strategy that has been called hierarchical shotgun sequencing. Genomic libraries are made in large insert BAC vectors. The clones from these libraries are arrayed and then subcloned into plasmid vectors. The resulting mini libraries are then randomly sequenced until enough sequence is obtained at high accuracy to assemble this shotgun sequence. Even the early stages of this assembly process are useful. So that investigators can have access to these incomplete or draft sequences, GenBank established the High Throughput Genome sequence (HTG) division. Within the draft or HTG sequences, GenBank recognizes different phases of completion. These phases are based on the degree of coverage and assembly of the sequence in the record. Phase 1 records have sufficient coverage to have several assembled regions (contigs). However the order and orientation of these is unknown, and there are still be gaps of unknown length in the sequence. Phase 2 records have progressed to the point that the order and orientation of the assemblies is known, but there are still gaps. As more sequence becomes available, the submitters update the records and the records will progress through the draft phases until the coverage and accuracy are sufficient for the sequence to move to phase 3. The record then will move from the HTG division into one of the traditional GenBank divisions: A human sequence then would move to the PRI division. A fly sequence would move to the INV division. A zebrafish sequence would move to the VRT division. Even though the sequences within them are incomplete, the draft sequences are still useful. The NCBI assembly of the human genome depends on draft sequences; as of July 2001, about half of the human genome is still in the HTG division.

The Sequence Tagged Site (STS) Division

STS division records are mapping reagents. A sequence tagged site is essentially a recipe for amplifying a specific fragment of genomic DNA using the polymerase chain reaction (PCR). The records generally include a pair of primers and the sequence of genomic DNA they amplify. STS markers are designed based on the sequence of a known gene, an EST, an mRNA or genetic marker. These markers are commonly used in the technique of radiation hybrid (RH) mapping as a means of constructing a physical map of a genomic region. In RH mapping a cell line from the species of interest (human for example) is given a lethal does of radiation. One effect of this is to break the genome into fragments. The fragmented genomic DNA of the irradiated cells can be rescued by fusing the irradiated cells with those of a different species. The resulting hybrid cells variously retain and expel fragments of the foreign genome so that unique clones from the hybrid line have differing portions of the foreign genome. Genomic DNA isolated from these clones can then be tested by PCR with STS markers to the irradiated genome. The

probability that markers occur together in the same hybrid clone is inversely related to the distance between them in the original genome. The pattern of amplification can thus be used to construct a physical map showing the relative positions of these markers. Since the genetic position of many of these markers is also known, radiation hybrid map positions can be integrated with genetic maps. Finally since STS markers are also sequenced based markers they can be mapped onto the assembled genomic sequence. The NCBI tool electronic PCR (ePCR) will search a sequence for the presence of markers from the STS division. This tool has been important in assembling the human genome sequence.

Other Special Divisions

The patent division (PAT) contains sequences provided by the US Patent and Trademark office. These sequences are not well annotated and may not be useful for patent claim investigation since GenBank cannot guarantee that this division includes all patents.

The contig division (CON) contains records that are instruction sets for assembling larger sequences. This division exists partly because GenBank has a 350 Kb limit for a single sequence. An example of a record in the CON division is the one containing instructions for assembling the *Escherichia coli* K12 genome from the < 350 Kb pieces in the BCT division.

The high throughput cDNA (HTC) division was recently created for draft cDNA records. Like the HTG division these sequences can be finished and then will move into the corresponding traditional division.

The Derivative Databases

Derivative databases use data from primary database like GenBank and add value by performing some kind of computational analysis or additional annotation and curation.

Protein only databases

Currently protein only sequence databases like PIR,the Protein Information Resource (http://pir.georgetown.edu/)and SWISS-PROT (http://www.expasy.ch/sprot/)are essentially derivative database because the majority of protein sequences in them come from translations of nucleotide sequences. Both of these databases curate the protein sequences extensively and add additional annotations. These include comparing various examples of the protein sequences derived from primary sources. Both the SWISS-PROT data and PIR data are available at the NCBI as a part of the Entrez system.

NCBI Secondary Databases

UniGene

The UniGene database contains sequence similarity-based clusters of expressed sequences. Naturally, the richest sources of expressed sequences are the EST data. These data over represent the number of transcripts because highly expressed messages will be present many times within the data. The goal of the UniGene is reduce the EST data and to identify all transcripts for a particular organism. UniGene data sets are available for those organisms with substantial EST data. The UniGene collections are a rich source for gene discovery. Because EST libraries are tissue specific, UniGene data can be used as a resource for gene expression information. The NCBI Serial Analysis of Gene expression pages and the CGAP pages take advantage of this latter feature.

UniGene Build Procedure

Expressed sequences and coding regions from genes are clustered by sequence similarity. This is done in stages after removing mitochondrial, vector sequences and masking for repetitive elements. An important problem is cross clustering of sequences for different genes. Cross clustering can arise because the level of sequencing errors in ESTs may approach the level of sequence divergence of members of the same gene family. One way to avoid some of this is to focus on the 3' untranslated regions first, since these are less well conserved than coding. Then clone based edges are added; this means adding 5' reads from clones whose 3' ends have already been clustered. Although the UniGene Data sets have "value added" to the primary EST data, the databases are not truly curated but are built automatically.

LocusLink and the RefSeq Project

Because of the tremendous growth in primary sequence data and the archival nature of these datasets, it can be difficult to identify the best sequence for a gene and in some cases even to find the sequence of interest because of confusing nomenclature problems. The LocusLink database attempts to solve some of these problems by collecting relevant links to sequences and other data in NCBI data as well as some outside databases. Each gene is assigned a stable unique identifier (Locus ID) and titles of entries are assigned based on the relevant genome nomenclature committee guidelines, the Human Genome Nomenclature Committee for the human genome. These titles are also propagated to the UniGene database as well to standardize nomenclature for the clusters. LocusLink also tracks historical name for the genes. The current scope is fruit fly, human, mouse, rat, and zebrafish.

RefSeq mRNAs and Proteins

A project that is intimately related to LocusLink is the generation of curated reference mRNA and protein sequences (RefSeqs) for the genes for the LocusLink entries. Collaborators supply information about which sequence is an appropriate representative

for a gene. To generate a reference mRNA sequence, the best representative primary database sequence that has a full-length coding region is chosen. This record is used to create provisional RefSeq records. Essentially this provisional RefSeq is a copy of the of the database sequence but also includes several annotation enhancements: additional publications, aliases, LocusID number, MIM number, map information, and official gene symbol and name. These provisional RefSeqs are then subject to human review. This review process provides further enhancements to the RefSeq including extension of the using sequence data in other GenBank records, or the literature, correction of sequencing errors, addition of additional publications and a summary of gene function. The final product represents a review article about the mRNA or protein. RefSeqs are available through LocusLink and are included in the Entrez and BLAST databases. RefSeq mRNA and protein sequences have distinctive accession numbers; NM_ followed by six digits for mRNA and NP_ followed by six digits for proteins.

Other NCBI Reference Sequences

There are several other kinds of reference sequences that are generated by projects at the NCBI.

Model Transcripts and Proteins

Closely related to the NM_ and NP_ RefSeqs are the model transcript and protein sequences. At this point these are generated only for the human genome by aligning the RefSeq mRNA to the corresponding genomic region. The genomic sequence that aligns is then used to create a model transcript and its corresponding translation. In many cases these sequences, do not match exactly the RefSeq mRNAs. This could be caused by assembly problems, sequencing error or true polymorphisms. These model sequences have distinctive accession numbers beginning with XM_ and XP_ and like the NM_ and NP_ RefSeqs are available through LocusLink, Entrez and through the BLAST databases.

NCBI Assemblies

NCBI has created it's own assembly of the human genome project data. The assembly consists of sets of contigs that are in turn built from assembling overlapping draft (HTG) and finished human sequence from GenBank. These large records are available thorugh LocusLink, the Entrez system and can be searched as a BLAST database on the Human Genome BLAST page. Their distinctive accession numbers begin with NT_. The chromosome records (NC_) that so far are created for the simpler complete genomes at NCBI are another RefSeq assembly.

Reference Genomic Records

The final type of RefSeq is a reference genomic record (NG_). These are created to serve as fixed regions of the human genome assembly They are needed where the sequence and

placement of a region is well known but difficult or impossible to assemble automatically. An example is the beta globin cluster on chromosome 11.

A summary of RefSeq accessions is given below.

RefSeq Accession	Type of record
NM_, NP_	Reference mRNA, translation
XM_, XP_	Model Transcript, translation
NT_	contig
NC_	Reference Chromosome
NG_	Reference Genomic

The GenBank Record

Each of the GenBank data files (e.g. gbpri1.seq.gz) on the ftp site is a compressed text file that contains thousands of sequences. The format of the individual records in this file is commonly known as the GenBank flat file format. This format consist of a series line types with various data fields on each line. The specifications of the record format are described in detail in the GenBank release notes (ftp://ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt)

Using the example below we will examine the kind of information in each section of a typical GenBank record.

```
LOCUS       AF062069     3808 bp    mRNA            INV       02-MAR-2000
DEFINITION  Limulus polyphemus myosin III mRNA, complete cds.
ACCESSION   AF062069
VERSION     AF062069.2  GI:7144484
KEYWORDS    .
SOURCE      Atlantic horseshoe crab.
  ORGANISM  Limulus polyphemus
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata;
            Xiphosura; Limulidae; Limulus.
REFERENCE   1  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     A myosin III from Limulus eyes is a clock-regulated phosphoprotein
  JOURNAL   J. Neurosci. (1998) In press
REFERENCE   2  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
REFERENCE   3  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REMARK    Sequence update by submitter
COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
FEATURES             Location/Qualifiers
     source          1..3808
                     /organism="Limulus polyphemus"
                     /db_xref="taxon:6850"
                     /tissue_type="lateral eye"
     CDS             258..3302
                     /note="N-terminal protein kinase domain; C-terminal myosin
                     heavy chain head; substrate for PKA"
                     /codon_start=1
                     /product="myosin III"
                     /protein_id="AAC16332.2"
                     /db_xref="GI:7144485"
                     /translation="MEYKCISEHLPFETLPDPGDRFEVQELVGTGTYATVYSAIDKQA
                                          <sequence omitted for brevity
                     LKYYSEEYLSRIYETHIKKIVKVQAIARKYFVKVRQSKTKPH"
BASE COUNT     1201 a    689 c    782 g   1136 t
ORIGIN
        1 tcgacatctg tggtcgcttt ttttagtaat aaaaaattgt attatgacgt cctatctgtt
                               <sequence omitted for brevity>
     3781 aagatacagt aactagggaa aaaaaaaa
//
```

LOCUS

The LOCUS line contains five fields: a locus name, molecule type, GenBank division code and modification date.

Originally the locus name was intended to be a unique identifier that also provided a meaningful name for a sequence. Early locus names from GenBank still reflect this intention; for example, MUSGAPDH for mouse glyceraldehyde 3-phosphate dehydrogenase or DROPROOX for *Drosophila melanogaster* proline oxidase. However as the number of organisms and genes represented increased, it became increasingly difficult to assign a Locus Name that was not merely another arbitrary identifier. For modern GenBank records we substitute the accession number in this position. You will notice that the EMBL database still assigns separate locus names to its records.

The next item in the LOCUS line is the sequence length. There is nothing particularly remarkable about this except that all lengths are in base pairs as for double stranded DNA. The molecule type for this record is mRNA. Of course the actual molecule sequenced was a cDNA. In fact all molecules represented in GenBank records are DNA. Other common molecule types are

DNA = genomic DNA
RNA = genomic RNA (RNA viruses)
rRNA = ribosomal RNA.

The fourth LOCUS line item is the three-letter GenBank division code. This record is from the invertebrate (INV) division, one of the traditional GenBank sequence divisions.

The last item in the LOCUS line is the modification date. Many people confuse this with the date of release of a record. In some cases, the modification date corresponds to the first release date of a sequence. However this is only true for records that have not been modified since their release. The modification date records the last time a record was changed by someone at GenBank. This could be something as trivial as correcting a typographical error or could be a large change in the sequence. Those interested in patenting sequences often need to know the first date of public release. This cannot be reliably determined from the GenBank record. Upon request we will research the release history of a record.

ACCESSION

The accession number is the main identifier for a GenBank record. It is the identifier that should be used when referring to a record in a publication. The accession number is intended to be a stable identifier and will not change when the sequence content is updated with additional or corrected data. More than one accession number may be in this section in some cases. This can happen when sequences are merged. Examples are when a larger sequence completely overlaps previous submissions from the same workers or when duplicate submissions are discovered. The first accession number represented is

then the primary accession. The other secondary accession numbers are left on the record so that a sequence may still be retrieved with previously published accession numbers. In many cases, the records corresponding to the secondary accessions will no longer be available separately.

VERSION

The version line tracks the revision history of a sequence. The version number is given as accession.version. The present example (AF062069.2) then is the second version of the sequence shown here. Unlike the modification date on the LOCUS line, which changes when any part of the record is altered, the version number will only change if there has been a change to the DNA sequence. Version numbers can go quite high for draft genome sequences, which may be updated more than twenty times before they are finished. There is a second identifier on the VERSION line known as the "GI" (gee – eye) number. The GI number is an internal database identifier that is unique to a particular version of a sequence. "GI", by the way, stands for GenInfo a pre-Entrez NCBI database system

DEFINITION

The DEFINITION line is the title of the GenBank record. The information on the DEFINITION line is indexed in the Title Word field of the Entrez search system. This line is intended to convey as much information as possible about the record and should give the source, the product and whether is a complete or partial sequence of that product. Our submission software usually constructs this line automatically.

KEYWORDS

The KEYWORDS line is deprecated at GenBank. We discourage the addition of keywords by submitters - our submission tool Sequin will not allow words to be added to this line. We limit keywords because of an unfortunate tendency to add large numbers of keywords that rightfully belong in some other part of the record or are redundant with information already present. The main exception is that we will add keywords for any bulk sequence division to this line. EST, GSS and STS division records will have the division code mirrored on this line. HTG division records will also have the phase indicated:

```
KEYWORDS    HTG; HTGS_PHASE1;
```

The HTG keyword will persist on finished genome project records that have moved into the appropriate traditional division of GenBank.

SOURCE, ORGANISM

The SOURCE line lists the accepted common name used in the NCBI taxonomy for the source of the sequence. The ORGANISM section provides the binomial scientific name of the organism followed by the taxonomic classification and lineage of the organism according to the NCBI taxonomy. We maintain a taxonomy database for all organisms for which we have sequence data.

REFERENCE

Usually a minimum of two REFERENCE blocks will appear on GenBank records. The first will always be something that at least is formatted like a literature citation. This is true even if the data will never be published in a journal. In some cases, there may be more than one primary literature citation. But the literature citations on GenBank records should only include those that report the primary sequence. The second kind of REFERENCE indicates who submitted the record. Finally if the record has been updated the will be another REFERENCE block that indicates the revision history.

COMMENT

This is a free text section of the records and just about anything can be here. When there has been an update the GI number of the previous sequence version will be given as in the current example. Only the most recent version of the sequence is included in the GenBank distribution on the ftp site. However in the Entrez system all versions of a sequence can be retrieved through their GI numbers. (To see an example of a very long COMMENT, view the GenBank record for the human beta globin region (U121317). This record also shows a very large number of secondary accession numbers.)

FEATURES

The section of the record between the header information and the sequence is the feature table. This is often the most complex and interesting part of the record. It contains a mapping of various kind biological data onto the primary sequence coordinate system. Feature tables can be enormous for large genomic records that have many genes, and repeat regions annotated. The feature tables may contain only the required source feature key for sequences from one of bulk sequence divisions (HTG, EST). The feature table for the present example is typical for an mRNA sequence from one of the traditional GenBank divisions. It has a source and CDS feature keys. The source key is required on all GenBank entries and indicates the biological source of the specified span. At a minimum there will be a source key that spans the entire length of the sequence with the mandatory “/organism” qualifier. The CDS feature key indicates a coding sequence. A CDS is a span of nucleotides that corresponds to the included protein translation. The span includes the first nucleotide of the translation start and the last nucleotide of the stop codon. There are a number of qualifiers associated with the CDS.

Two of them report the corresponding protein accession and GI numbers for the translation

```
/protein_id="AAC16332.2"
/db_xref="GI:7144485"
```

Recall that there are no protein sequences in GenBank proper. We do make these translations available as a separate GenPept data set to which these identifiers apply. Of course these translations are available as a part of the Entrez protein database. Notice that, from the GenBank point of view, this implied protein is a feature of the DNA sequence. This is reasonable since, in many cases, there is no real evidence of the protein other than the open reading frame on the mRNA sequence or the predicted gene on a genomic record. The DDBJ/EMBL/GenBank Feature Table Documentation (http://www.ncbi.nlm.nih.gov/collab/FT/index.html) provides detailed information on the kinds of features that are annotated on GenBank records.

ORIGIN

The DNA sequence in lower case single letters follows the immediately following the ORIGIN. The sequence is grouped in blocks of ten nucleotides, 60 per line. The end of the record is indicated by the double front slash '//'. In the ftp site sequence files another record would immediately follow.